

RE-RecSys: An End-to-End system for recommending properties in Real-Estate domain

Venkatesh C

Housing.com, India

venkatesh.c@housing.com

Anil Goyal

Housing.com, India

anil.goyal@housing.com

Harshit Oberoi

Housing.com, India

harshit.oberoi@housing.com

Nikhil Sikka

Housing.com, India

nikhil.sikka@housing.com

ABSTRACT

We propose an end-to-end real-estate recommendation system, RE-RecSys, which has been productionized in real-world industry setting. We categorize any user into 4 categories based on available historical data: *i*) cold-start users; *ii*) short-term users; *iii*) long-term users; and *iv*) short-long term users. For cold-start users, we propose a novel rule-based engine that is based on the popularity of locality and user preferences. For short-term users, we propose to use content-filtering model which recommends properties based on recent interactions of users. For long-term and short-long term users, we propose a novel combination of content and collaborative filtering based approach which can be easily productionized in the real-world scenario. Moreover, based on the conversion rate, we have designed a novel weighing scheme for different impressions done by users on the platform for the training of content and collaborative models. Finally, we show the efficiency of the proposed pipeline, RE-RecSys, on a real-world property and click stream dataset collected from leading real-estate platform in India. We show that the proposed pipeline is deployable in real-world scenario with an average latency of <40 ms serving 1000 rpm.

CCS CONCEPTS

- Computing methodologies → Machine learning algorithms.

KEYWORDS

recommendation engines, PDP widgets, matrix factorization, cold start

ACM Reference Format:

Venkatesh C, Harshit Oberoi, Anil Goyal, and Nikhil Sikka. 2024. RE-RecSys: An End-to-End system for recommending properties in Real-Estate domain. In *7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD) (CODS-COMAD 2024), January 4–7, 2024, Bangalore, India*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3632410.3632487>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CODS-COMAD 2024, January 4–7, 2024, Bangalore, India

© 2024 Association for Computing Machinery.

ACM ISBN 979-8-4007-1634-8/24/01...\$15.00

<https://doi.org/10.1145/3632410.3632487>

1 INTRODUCTION

Over the past few years, the demand for online real-estate tools has increased drastically due to the ease of access to the Internet, especially in developing countries like India. There are many online real-estate platforms for owners, developers, and real-estate brokers to post properties for buying and renting purposes. These platforms have more than 3 million active customers per month and have more than 1.1 million active properties. Considering the number of users and properties, it makes it difficult for users to find relevant properties from a large list of possibilities. Recommendation Systems (RS) are used to provide personalized recommendations based on historical interactions with the platform. The development of an efficient RS is critical from both the company's and the customer's point of view. On the one hand, it helps customers to narrow down their choices leading to higher customer satisfaction and retention. On the other hand, it helps companies to increase their traffic, conversions, and click-through rates (CTRs). In this paper, we focus on the real estate recommendation engines where housing selection (for both rent and purchase) is a complex decision-making procedure because people purchase/rent properties infrequently throughout their life. Typically, users come on an online platform via web/mobile application and expresses their needs using search criteria such as rent/purchase, location, price, locality, number of rooms, etc. Moreover, they interact with the listed properties on the platform by clicking the property details page (PDP), viewing images, playing property videos, filling customer requirement forms (CRF), dropping leads, etc. All these interactions on the platform contribute to user preferences and help RS capture the user preferences for providing relevant recommendations.

In this work, we proposed an end-to-end pipeline for **real-estate recommendation system** (called as RE-RecSys) which we have productionized in real-world industry setting. Concretely, we classify any user into 4 categories: *i*) *cold-start* users with no historical data; *ii*) *short-term* users with recent 10 minutes of interactions on the platform; *iii*) *long-term* who have more than 10 minutes of historical data and interacted with at-least 5 properties on the platform; and *iv*) *short-long term* users who have both historical data as well as recent interactions on the platform. As we are in an industry setting, it is very crucial to have low latency (< 40 ms serving 1000 requests per minute) to have a good customer experience on the platform. Therefore, we propose to use a combination of rule-based, content, and collaborative filtering for the real-estate recommendation system. Specifically, for *cold-start* users, once users express their needs in search criteria, we recommend those properties with the highest

number of leads, user conversions, and property conversions in the chosen locality of a particular city. For *short-term* users, we propose to use content-based filtering [1, 8] which makes use of existing contextual information about the users (e.g. location, price, apartment type) and properties (e.g. apartment details) for a recommendation. For *long-term* users, we used combination of content and collaborative filtering [3] which relies on past interactions and recommends properties to users based on the interactions done by other similar users. For *short-long term* users, we used a combination of short-term and long-term models. We answer the following research questions for real-estate recommendation engines:

- (1) What should be the ideal amount of historical training data required for rent and purchase purposes?
- (2) What user impressions/interactions should be considered as implicit feedback for both content and collaborative filtering?
- (3) How to deploy the solution in a production environment keeping the latency < 40 ms?¹

2 PROPOSED APPROACH

Purchasing/renting houses is a complex process as they are expensive and people usually purchase/rent them infrequently. Moreover, the behavior and preferences of customers change over time (due to property visits, budget constraints, etc) which adds additional temporal complexity to the housing recommendation problem statement. It is important to consider the recent as well as past history of a user while designing a recommendation engine for real estate. Therefore, we propose to classify users based on the availability of historical data: cold-start users, short-term users, long-term users, and short-long term users. In the next subsections, we will explain how we tackled each kind of user and built an end-to-end pipeline.

2.0.1 Rule-Based engine for Cold-Start Users. On real-estate platforms, approximately 25% of users are cold-start users and we have no available historical data for model training. As users' short/long history is not available on the platform therefore content and collaborative filtering based methods would not be applicable [9]. Therefore, we propose to use a rule engine to handle cold start users on the platform and keep the latency of the model in acceptable limits for the production scenarios.

Firstly, for each locality in the city, we build cohorts (in other words, groups), based on locality name, apartment type (e.g. 2 BHK, 3, BHK, etc), profile type (broker and owner), price per square feet bins (bins are created with a gap of 500K and 10k Rupees for purchase and rent respectively) and area bins (created with a gap of 500 sq feet area). Due to this, we are able to narrow down the search space for a particular search query based on filters. Then, we calculate the score for each cohort based on the summation of the following 4 metrics: *number of flats* (indicative of density); *total leads* (indicative of popularity); *percentage of property conversions* (another indicative of popularity); and *percentage of user conversions* (indicative of user preference). Please note that the last 2 metrics inherently take care of scenarios where the number of properties in a particular cohort may not be high. Finally, the top-*N* matching cohorts are extracted based on search filters for any user, then we

return the randomly chosen top 2 properties from each cohort. Intuitively, the random selection of properties from each cohort helps us to overcome the challenge of cold-start problem for new properties. Each new property will fall under some cohort, randomly choosing properties from each cohort allow us to increase the reach for them which will lead to better customer satisfaction. Moreover, it helps to have better learning for content and collaborative filtering models.

2.0.2 Content Filtering for Short-term Users. Approximately 20% of users fall under the category of short-term users where we have recent 10 minutes of interaction data on the platform. Therefore, it is important to personalize the experience for this category of users which can lead to better conversions and CTRs. One possible solution is to use collaborative filtering which relies on past interactions and recommend properties based on interactions done by other similar users. However, given the number of users and properties, it would not be possible to re-train the collaborative filtering model at the regular interval of 10 minutes [2, 4]. Thus, we propose to use content-based filtering approach for short-term users in order to have a better personalization experience in real-time scenario. We calculate the similarity scores between the user and properties using a cosine similarity measure where users and properties are represented in the same feature space. Then, we rank the properties based on the calculated similarity scores to return the results for a particular search query filter.

For the property vector, we have considered multiple categorical (apartment type and furnishing type) and numerical (price, built-up area, age, floor number and image count) features based on the property description. We convert each feature into a categorical feature and create a binary vector for the property vector. For area, age, floor number, and image count, we created bins with a distance of 500 sq feet, 3 years of age, 2 floors, and 3 images, respectively. For the price, we created bins with a gap of 500k and 10k for purchase and rent properties respectively. The apartment type (2 BHK, 3 BHK, etc.) and furnishing type (fully furnished, unfurnished, and semi-furnished) are categorical variables kept in their original form.

Users on the platform have different levels of behavior such as clicking, viewing the property details page, dropping a lead using CRF forms, viewing images, etc. It is important that different behaviors should receive different weights in the user profile vector. We have divided different activities into 4 categories based on actions: *i*) “conversion” which take into account all the activities where user is planning to submit a lead by filling customer requirement form (CRF); *ii*) “detail page” where user is viewing the details for a particular property after opening product detail page; *iii*) “impressions” where the user is exploring property by looking into its miscellaneous details; and *iv*) “other” activities where user is just scrolling the page and spending some time on rating and other details. For any user activity, we calculated the conversion rates from that activity till submitting the CRF form. Based on the conversion rates, we have assigned weights to each action as shown in Table 1. We have assigned the highest weight to the activity which has the highest conversion rate and maximum business impact. Then, we calculate the user profile vector similar to the property vector by converting each feature into the categorical feature as explained previously. For the user profile vector, we calculate the weighted sum of all the activities done by the user over various properties instead of binary

¹Demo Video is available at <https://www.youtube.com/watch?v=On2JGxACnag>

Table 1: Assigned Weightages for different actions performed by the user for both purchase and rent

Category	Actions	Conversion (Purchase)	Conversion (Rent)	Weight
Conversion	Submitted CRF	100	100	10
Conversion	One time Password	83.5	83.1	8
Conversion	Open or Filled CRF	36.8	36.9	6
Detail Page	recommendation, image/video views	26.1	27.7	4
Impressions	locality info, amenities check, price/floor plan	16.8	14.7	2
Other	Rating Check, open PDP & scrolling	23.8	19.6	1

encoding. This allows us to capture the user’s inclination towards a particular attribute. Based on the last 10 minutes of interactions for any user, we calculate the user profile vector and compare it with property vectors in a given locality using cosine similarity. Finally, we return the top- N most similar properties to the user based on similarity scores.

2.0.3 Collaborative Filtering for Long-term Users. On real-estate platforms, approximately 35% of users have historical data of more than 10 minutes and have interactions with at least 5 properties on the platform. For this kind of users, we propose to use collaborative filtering which recommends properties based on the preferences of other similar users. Since, in the case of real-estate recommendation problem, user-property matrix is highly sparse and it is important to recommend less popular properties as well as well-known properties. Therefore, we propose to use the matrix factorization technique which factorizes a user-property matrix into two low-ranked matrices, the user-factor matrix and the item-factor matrix that can predict new items for a particular user [6, 7]. For each user-property pair in the matrix, we will compute the score based on the interaction done by a user for a particular property as defined in Table 1. For e.g. for a particular property if a user has submitted the CRF form then we fill the score with value of 10 or if user have just scrolled the property page then we fill the score with the value of 1. This allows us to capture the user preferences at the interaction level and provide more personalized experience. For matrix factorization, we use Alternating Least Square algorithm[5] which is implemented in Apache Spark ML and built for large-scale collaborative filtering problems. The training time for ALS is approximately 30 minutes for 3 million and 1.1 million users and properties respectively. Therefore, it is possible to re-train the collaborative model at regular intervals of 1-2 hours. However, for collaborative filtering, there would not be any significant performance improvement with 1-2 hours of additional data. Therefore, we propose to re-train the collaborative filtering model once a day and use content filtering for users who have interacted with more than five properties but were not included in the previous ALS re-training. Please note that the content model for long-term users is updated at regular intervals of 2 hours. Compared to the content model for short-term users (updated at every 10 minutes intervals), long-term content model have bigger time complexity due to large number of interactions for long-term users. In production, we have deployed a content+collaborative model for long-term users.

2.0.4 Hybrid approach for Short-long term Users. In recommendation engine pipelines, it is common to have users with both

recent as well as past history on the platform. On real-estate platforms, approximately 20% of users fall under this category. One solution is to use long-term collaborative filtering model to recommend properties to users. However, long-term models are re-trained with a gap of regular intervals of 24 hours and don’t take into account the recent history of the user. Therefore, we propose to use a hybrid approach which is a weighted combination of content and collaborative filtering models. For any user, we take the average of scores from both models and return the top properties accordingly.

2.0.5 RE-RecSys Pipeline. As shown in figure 1, for any new user, the first 10 minutes are served using the cold-start rule-based engine. If the user is active for 2 hours and have interacted with more than 5 flats, then the user will be served with content (re-trained every 2 hours)+collaborative model (re-trained every 24 hours) model. If the user is active for 2 – 4 hours, then hybrid model (long-term and short-term models) will return the results to take into account the recent history of user. In case of user inactivity for more than 28 days then the user will be served using only cold-start model.

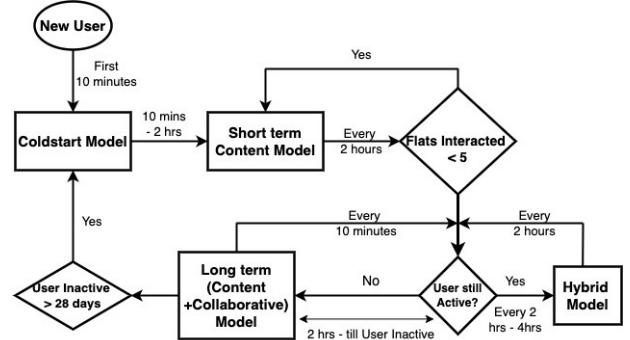


Figure 1: An End-to-End architecture diagram for RE-RecSys

3 EXPERIMENTAL RESULTS AND DEMO

3.0.1 Dataset. We have collected the real-world dataset from our platform which is a leading real-estate online platform in India. The platform has an average of 3 million active users per month and 1.1 million active properties. Moreover, users view 3.57 pages per visit on average indicating high engagement and interest in the platform. For the real-estate recommendation system, we need 2 set of information: *i*) property-related data and; *ii*) user interaction events on the platform. For the property dataset, we used an internal relational database to collect the information related to properties such as location, price, area of the apartment, property type, etc. We use Google Analytics to track the user interaction events (1+ billion clicks per month) on our platform. For our analysis, we have collected the 6 months of property and event datasets from 1st January to 30th June 2022.

3.0.2 Life Cycle of Buy and Rent Users. As we are in a real-world scenario where users rent/purchase houses infrequently. So, it is important to understand the time taken by the user to make a final decision on the platform. Moreover, from a machine learning point of view, this analysis is crucial for setting up the training and testing pipelines. From the collected dataset, we analyzed the user persistence in terms of days for both type of users. From our

experiments, we validated that 85% of rent and buy users persist for approximately 3 months (95 days) and 6 months (190 days) on the platform. Therefore, we used the latest 3 and 6 months of data for rent and buy users respectively for re-training our models.

3.0.3 Train-Test Split. For collected 6 months of data, we have randomly sampled click event data consisting of 20,000 users for each kind of task (cold-start, short-term, long-term and short-long users) separately for rent and purchase use cases. In total, we collected 160,000 user data. Each user's history is split into train and test data based on random checkpoints. By doing this, we are able to test our models on various kinds of users having different amounts of historical data, preferences, and interactions with the platform. Therefore, all the models (content or collaborative) are trained with 20,000 of training data and are tested with 20,000 of test data separately for rent and buy. As followed in literature [1, 6], we used MAP@K (Mean Average Precision at K) and NDCG (Normalized Discounted Cumulative Gain) as our evaluation metrics.

3.0.4 Experimental Results. Firstly, we evaluated the performance of the content-filtering model on short-term users. Here our objective is to analyze the amount of data required for training the model. In table 2, we present the results for different iterations of the latest training data i.e. 5, 10, 20, and 30 minutes. It means we train the model with the last x minutes of data and test it on the next x minutes. From the table, we can deduce that the best results are achieved when we train the model with last 5 minutes of data. However, in the production environment, for training the content model, we need to extract the real-time click event data from Google Analytics, pre-process and compute the features. This is a time taking process and could not be finished within 5 minutes given the number of users on the platform. Therefore, in production, we used the latest 10 minutes of data for training the content model for short-term users. Moreover, the average inference latency of content model for short-term user is 23.1 ms (1000 requests per minute) which is in acceptable limit in production.

Table 2: Results for content filtering on short-term users

Experiment	Buy		Rent	
Train & Test Sets	MAP@6	NDCG	MAP@6	NDCG
latest & next 30 min	0.853	0.662	0.839	0.625
latest & next 20 min	0.856	0.664	0.845	0.629
latest & next 10 min	0.866	0.673	0.856	0.636
latest & next 5 min	0.881	0.69	0.873	0.646

Secondly, we have validated the proposed collaborative-filtering model for long-term users trained using the alternating least square algorithm [5]. We have compared the proposed approach (linear weighing of click events) with two other approaches:

- Exponential Decay:** We use half-life exponential decay for event weights at regular intervals of 3 days. Here, we give more weight to activities that are recent as compared to old impressions.
- TF-IDF Weighing:** We multiply the event weight with inverse property frequency defined as $\log(\frac{\text{total_interactions}}{\text{interactions_on_property}})$. We give more weightage to less popular properties in the corpus.

From results in table 3, we can deduce that linear weighing performs best in terms of NDCG for buy and in terms of MAP@6 for rent. For other cases, it is second best compared to baselines. In the

production environment, the average inference latency for linear weighing collaborative model is 19.29 ms (1000 requests per minute). Finally, for *short-long* term users, we use a hybrid approach where we take the average of scores from both models (long-term and short-term) and return the top properties accordingly. Moreover, we have evaluated the inference latency for this model and it is within our acceptable limits i.e. 29.3 ms (1000 requests per minute).

Table 3: Results for collaborative filtering for long-term users

Experiment	Buy		Rent	
	MAP@6	NDCG	MAP@6	NDCG
TF-IDF weighting	0.713	0.655	0.614	0.691
Exponential decay	0.865	0.662	0.65	0.596
Linear weighting	0.823	0.685	0.804	0.646

3.0.5 REST API and Integration with UI. RE-RecSys system is developed in Python and released as REST API. In figure 2, we present an example of internal API call along with a JSON response. Finally, the results (property ids and corresponding scores) obtained from the API are integrated with various front-end widgets on mobile application, as shown in the figure 3.²

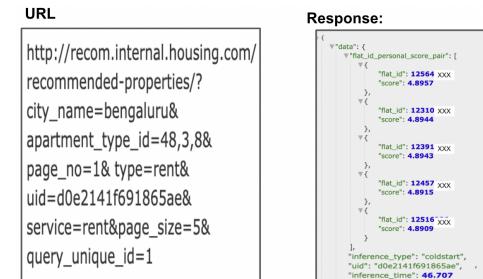


Figure 2: Demo of internal RE-RecSys call Panel

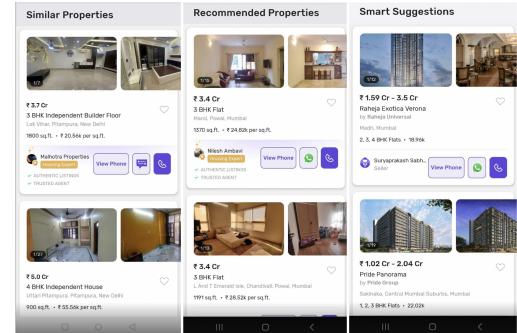


Figure 3: RE-RecSys integrated with Housing.com App

4 CONCLUSION

We propose an end-to-end pipeline, RE-RecSys, for recommending properties to real-estate users. We designed an architecture which can handle different kind of users (cold-start, short-term, long-term, and short-long term users) keeping a balance of infrastructure costs and near real-time personalized recommendations (< 40ms serving 1000 rpm). We have evaluated the performance of proposed algorithms on a sub-sample of data consisting of interactions from 160,000 users and compared it with baselines.

²Demo Video is available at <https://www.youtube.com/watch?v=On2JGxACnag>

REFERENCES

- [1] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013), 109–132.
- [2] Sabri Boutemedjet and Djemel Ziou. 2012. Predictive approach for user long-term needs in content-based image suggestion. *IEEE transactions on neural networks and learning systems* 23, 8 (2012), 1242–1253.
- [3] Rui Chen, Qingyi Hua, Yan-Shuo Chang, Bo Wang, Lei Zhang, and Xiangjie Kong. 2018. A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. *IEEE Access* 6 (2018), 64301–64320.
- [4] Yupeng Gu, Bo Zhao, David Hardtke, and Yizhou Sun. 2016. Learning global term weights for content-based recommender systems. In *Proceedings of the 25th International Conference on World Wide Web*. 391–400.
- [5] Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. 2015. Matrix completion and low-rank SVD via fast alternating least squares. *The Journal of Machine Learning Research* 16, 1 (2015), 3367–3402.
- [6] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender systems: an introduction*. Cambridge University Press.
- [7] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
- [8] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. *Recommender systems handbook* (2011), 73–105.
- [9] Faiza Rehman, Hira Masood, Adnan Ul-Hasan, Raheel Nawaz, and Faisal Shafait. 2020. An intelligent context aware recommender system for real-estate. In *Pattern Recognition and Artificial Intelligence: Third Mediterranean Conference, MedPRAI 2019, Istanbul, Turkey, December 22–23, 2019, Proceedings* 3. Springer, 177–191.